Editorial:

Editorial: Avoid being tripped up by statistics: Statistical guidance for a successful research paper.

Robin John Prescott

Centre for Population Health Sciences, Usher Institute, University of Edinburgh, UK
robin.prescott@ed.ac.uk

Declarations of interests: none

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Editorial: Avoid being tripped up by statistics: Statistical guidance for a successful research paper.

Introduction

More than 40 years ago it was recognised that a high proportion of papers published in medical journals contained statistical errors [1]. There have been countless textbooks and educational papers in medical journals addressing this problem. Journals have increasingly used statistical reviewers for at least some submissions and some of the major journals use statistical reviewers for every submission with any statistical content. This has undoubtedly improved the quality of published articles to some extent but problems still persist [2]. The end product that appears in journals has often benefitted through the statistical input of reviewers and weaknesses in the original submissions are not always apparent. In a survey of papers submitted to *Injury,* 90% needed some form of correction after statistical review [3].

This article is directed at improving the statistical quality of papers submitted to this journal and, hopefully, others. It is not attempting to do the impossible and cover all statistical issues. Rather, it focuses on issues that have recurred repeatedly in the author's experience of reviewing around 2,000 submissions to medical journals in recent years. In this paper, it will be assumed that the reader has a basic knowledge of statistics and understands concepts such as tests of significance and confidence intervals. For the complete novice, there are dozens of introductory textbooks on medical statistics and a great deal of help is available online. I always hesitate to recommend a specific text or source as teaching experience showed me that some students have their eyes opened by one text while another finds the same source obscure but likes an alternative. Generally popular introductory books were those by Martin Bland [4] or Kirkwood and Sterne [5], while Altman [6] goes beyond the basics in his text.

Some of the recommendations in this paper will be controversial. The reader should be aware that repeated use of a statistical technique in the literature is not a guarantee of its validity and, in such instances, this paper will try to explain the basis of the criticism without going into details of statistical theory. The focus is on encouraging good practice in areas where others have had difficulty or have, inadvertently, been using inferior methods.

Reporting Guidelines

There has been a substantial increase in the number of available reporting guidelines in recent years. This has stemmed from a perceived weakness in many different research areas of how research has been conducted and presented. Historically there has been a frequent lack of clarity in many publications and/or poor quality and the earliest attempts to address this problem were the construction of check lists which were initially designed to be used by readers evaluating published research, or by reviewers. These have been built upon over time and have led to the development of guidelines in specific areas to ensure that the methods used when conducting the research are appropriate, that the important elements of the design are reported and that there is clarity in the presentation of results. CONSORT [7], for the reporting of clinical trials, was one of the earliest and perhaps the most widely known guideline, first published in 1996. In common with many guidelines it has been refined with successive updates to further improve communication.

Some of the more widely used guidelines and their area of application are shown in Table 1. Equator network (www.equator-network.org) provides a comprehensive searchable database of reporting guidelines for health research (389 of them at the time of writing) and it is well worth checking to see if there is a guideline that will help with one's own research. For example, among the less well-known, there are recommendations for the reporting of foot and ankle models [8]. As a statistician I would particularly encourage use of the SAMPL guidelines [9]. These cover basic

statistical reporting for articles published in biomedical journals. In the course of reviewing, I have had occasion to direct authors to this guideline on many occasions.

If there is a relevant guideline I would urge authors to use it. I have no doubt that they have improved the quality of publications and reviewers will largely expect those available to be used. If authors decide a guideline that could be relevant is inappropriate for their purposes, it is sensible for them to 'get their retaliation in first' and describe within Methods why a reporting guideline has not been used.

Missing Values

Missing values occur in many studies for a variety of reasons. Although the ideal is to have complete data as that simplifies both analysis and presentation, the reality is that missing values are common. There is a temptation, when writing up results, to simply ignore the occurrence of missing values and report an analysis based on the observations that are available. This is unacceptable on many fronts. Firstly, it means that the reporting is less than transparent, which is always to be avoided. Inconvenient facts should never be swept under the carpet. Secondly, the presence of missing observations may cause bias in the results that are presented and it is important that the reader is aware of this possibility. Thirdly, ignoring the existence of missing data negates the possibility of using modern statistical methods such as multiple imputation to mitigate the effect of missing values.

As well as the guidelines mentioned in the previous section, any author with missing values would also be well advised to read an excellent paper by Sterne and colleagues [10]. This includes guidelines for reporting any analysis potentially affected by missing data. It also provides an introduction to the types of missing data and a discussion of methods to handle missing data. Multiple imputation is introduced along with its benefits and pitfalls.

If a study has been performed in such a way that missing values have been avoided, a positive statement to that effect should always be included to save the reader having to second-guess whether or not there were missing values.

Parametric and Non-parametric Tests and Testing for Normality

Many statistical tests of significance are based on an assumption of Normal distributions (parametric methods). For example, in an unpaired t-test, we assume that each sample is drawn from a population in which the variable is normally distributed. In a standard regression model or in an analysis of variance (ANOVA), we assume that the residuals are normally distributed. It is therefore logical to consider the use of statistical tests to determine if that assumption of normality is valid and many papers published in this journal have used a Kolmogorov-Smirnov test or the Shapiro-Wilk test, among others. That choice is likely to be strongly influenced by the choices offered in the statistical software package used.

Although such testing appears logical, there are good reasons to avoid these tests. Parametric tests have long been recognised to be robust to deviations from the assumption of normality. This derives from the Central Limit Theorem which states that, as sample sizes become large enough, sample means are normally distributed, whatever the distribution of the variable in the population from which the sample is drawn. An illustration of this remarkable result comes from the early days of computing. In order to obtain a random sample from a normal distribution, twelve samples were taken from a rectangular distribution (which was easy to sample) and their values added. The agreement to normality was very close except in the extreme tails of the distribution.

The only situations where the robustness of parametric methods is threatened is when samples are very small or when there is a modest sample size with extreme skewness in the variable of interest. If a test for normality is performed when the sample sizes are small, there is the problem

that the tests have extremely low power. That is, the tests are unlikely to detect non-normality, even if it exists and we may be misleadingly reassured that the normality assumption is sound. With large sample sizes, the tests for normality will be able to detect small departures from normality, that would not invalidate the parametric methods. Thus, the benefit from testing for normality is illusory.

In assessing whether parametric methods are justified, it is the skewness of the data that is most important. This can be assessed graphically in several ways including using a simple dot plot or a more sophisticated Q-Q plot, where normally distributed samples will lie on a straight line. These will be available on most statistical software packages. A coefficient of skewness can also be calculated, but the graphical methods are perhaps more directly appreciated. There are no simple rules that can be applied but it is only in smaller samples where the decision is likely to be critical. With a sample size in single figures I would examine the distributions with great care. Once sample sizes are above 30, only the most extreme skewness would be of concern to me. In between, moderate to strong skewness could be a worrying issue.

When using analysis of variance or regression methods, it is important to recognise that it is only the normality of residuals that matters and non-normality in the dependent variable will not necessarily mean that residuals are non-normally distributed. The more terms involved in the ANOVA or regression, the better the agreement to normality in the residuals tends to be.

If there are concerns about the use of parametric methods because of the observed distributions, one possibility is to transform the data. Positively skew data are widespread in medical applications and the use of a logarithmic transformation will often be successful in achieving approximate normality. Generalised linear models can also be considered [11]. An introduction to this topic can be readily obtained by a Google search. Otherwise, non-parametric methods may be necessary. There are drawbacks to using non-parametric methods however. There is not the flexibility in analysis that is offered by a more complicated analysis of variance or by regression models. There is also some loss of power in using non-parametric methods when the distribution is actually normal and this is most apparent with smaller sample sizes.

Another drawback to non-parametric methods has been illustrated by Fagerland [12]. He shows that with skewed data, with equal means and medians but with different dispersions, the Null hypothesis is rejected more often with the Wilcoxon-Mann-Whitney test than with the t-test. His conclusion is worth noting: “Non-parametric tests are most useful for small studies. Research authors that use non-parametric tests in large studies may provide answers to the wrong question, thus confusing readers. For large studies, t-tests and their corresponding confidence intervals can and should be used even for heavily skewed data”.

This author strongly supports these views of Fagerland who also advises against testing for normality. In the vast majority of cases, researchers should be analysing their data using parametric methods, possibly after a transformation if their data is skew. Non-parametric methods should only be considered as a last resort with small sample sizes and when it is apparent that a transformation is not able to normalise the data. Tests for normality should never be part of that process.

Tests of Significance, Confidence Intervals and Post-hoc Power

Statistical tests of significance are a feature of medical research papers that is rarely absent. They are a valuable way of assessing the plausibility of a Null Hypothesis and there is usually the hope that it will be rejected in favour of an alternative hypothesis. For example, in the comparison of two groups we will often be interested in rejecting the Null Hypothesis that both groups were sampled from the same population (e.g. with equal means) and concluding that there is a significantly higher mean value in one of the two groups compared to the other. As well as interest in the acceptance/rejection of the Null Hypothesis, we should also be interested in what magnitude of difference between the two groups is compatible with the data. We usually express this as a 95% confidence interval - a range within which there is a 95% probability that the true but unknown population difference will lie.

If there is a genuine clinically important difference between the two groups (say d) we would like there to be a high probability of correctly rejecting the Null Hypothesis. This probability is known as the power of the study and it differs according to the value of d. This is also equivalent to the probability that a 95% confidence interval will exclude a value of 0 when the true difference is d.

In some studies, when the research is reported, the question will arise as to whether the sample size was sufficiently large to have a high power of detecting important differences. One approach that is often advocated to answer this question is to calculate the post-hoc power. This is defined as the power the study would have to conclude that the observed difference was statistically significant. This is a deceptively attractive concept but it is one that has very little value, if any at all. The post-hoc power is a direct function of the observed p-value. In particular, if $p<0.05$, the post-hoc power will be greater than 50% and if $p>0.05$ the post-hoc power will be less than 50%. In essence, the post-hoc power is just presenting the p-value in an alternative format.

It will always be more informative to show the 95% confidence interval and form a conclusion on the adequacy of the sample size depending on whether the confidence interval gives a sufficiently precise answer for the clinical circumstances. If the answer is insufficiently precise it can be helpful to utilise the observed variability in the current study to make a formal sample size calculation. This will give the necessary sample size to have a specified power (often 80%) for the difference, d, that is deemed to be clinically important. Note that this does not use the observed difference but the difference that has clinical relevance. This is very different from the unhelpful but widely reported post-hoc power.

Effect Size

I have been asked to include a section on effect size as this is something reviewers sometimes request from authors. In comparing the effects of two treatments, if the main outcome variable is quantitative, the most obvious summary measure of the treatment effect is the difference between the means of the response variable in the two groups. To take a simple example, if we were measuring blood pressure, that difference would be measured in units of mmHg. With such a familiar and well understood measurement there would be no need for any other summary of the magnitude of the treatment difference. If a relatively unfamiliar outcome measure was in use, the magnitude of the treatment difference may be difficult to interpret. One answer to that dilemma is to standardise the outcome measurement. Thus we could, for example, compute Cohen's d, which is the difference between the means divided by the pooled standard deviation from the two groups. Cohen's d is a measure of effect size and it occurs widely in the literature. It is by no means the only effect size measure available and the closely related standardised mean difference (SMD) is another.

The SMD is widely used in meta analyses because of its endorsement by Cochrane (formerly called the Cochrane Collaboration). If different measurements have been taken in various studies but all are directed at a common concept their effects can each be summarised by a SMD and these can then be pooled. For example, balance capacity could be assessed in different ways in the individual studies within a meta analysis but they can be combined into a pooled effect by the use of SMDs.

Thus the use of effect sizes can be helpful but a caveat should be noted. The results will be affected by the magnitude of the pooled standard deviation. Thus effect sizes will tend to be larger in homogeneous populations with a smaller SD and smaller in heterogeneous populations, even if treatment effects are identical. Therefore if a common measurement has been used in studies contributing to a meta analysis it is preferable to use that measurement directly rather than using SMDs. In individual studies, if the outcome measurements are well known, there is no benefit in standardising them.

In summary effect sizes can be useful in the right circumstances but they should not be used automatically.

What is the Smallest Acceptable Sample Size?

The journal has received submissions which have validated new techniques on just one or two subjects. I was asked to consider what should be the lowest acceptable sample size for such a validation. There is no simple answer to that question as it is context dependent. For example, if validating a predictive model with a binary outcome, such as occurrence of a fall, several dozen would be needed to have even basic estimates of sensitivity and specificity. If, however, it was a continuous outcome that was being predicted, a very much smaller sample size would be acceptable.

Whenever measurements are being made, the sample size should be sufficient to permit at least a basic description of the distribution of the measurements and for 95% confidence intervals for the mean to be appropriately small. Assuming a normally distributed variable, the total length of the 95% confidence interval for the mean will be 17.97xSD when the sample size is 2, 4.97xSD when the sample size is 3, 3.18xSD with a sample size of 4, 2.48xSD with a sample size of 5, 2.10xSD with a sample size of 6, 1.85xSD with a sample size of 7 and 1.67xSD with sample size of 8. These examples show the considerable value of extra observations when the sample sizes are small. Additionally, the greater the number of observations, the easier it is to detect if the distribution is markedly skew, necessitating a transformation of the data.

In an ideal world, formal statistical sample size calculations would determine the size of every study but in the real world this is often impractical for sound reasons. My recommendation is that all studies should provide a justification of the sample sizes that they have employed, with the observation that it will always be increasingly difficult to justify sample sizes lower than 8.

Independence of Observations

All of the standard, basic, statistical techniques found in any introductory textbook on medical statistics are based on an assumption of independence. Even those techniques such as a paired t-test, which takes into account an inherent dependence between the members of each pair, still has the assumption that the ‘residual’ or ‘error’ terms are independent. In gait research, it is common for variables to be assessed for each individual over a number of gait cycles and/or over a number of repeated ‘trials’. Observations could be made on both the left and right legs. Under such circumstances, the individual observations cannot be utilised directly in a ‘standard’ statistical analysis. The effective sample size will be artificially increased if this is done. Often, means from these repeated observations are calculated so that a ‘standard’ analysis can be applied to the means from each individual.

An alternative approach is to use more advanced statistical methods that can take the correlations between these repeated observations into account in an appropriate way [13]. These are commonly described as mixed models or, equivalently, multi-level models. In studies where the degree of replication is identical for all participants, the use of means will still provide an efficient analysis. If there is any imbalance, however, the mixed model will always be more efficient (i.e. have greater statistical power to detect differences of interest).

For example, Stewart et al. [14] applied four methods of analysis to a dataset containing plantar pressure data which was collected for seven sites on right and left feet, over three trials, across three participant groups. The first three methods were ‘conventional’, with the data being averaged over trials. One method analysed right foot data, the second selected a random foot and the third averaged left and right foot data. The fourth method used a mixed model accounting for repeated measures for each foot, foot site and trial. The authors found that between group confidence interval widths were consistently smaller for the mixed models method and this also revealed more statistically significant between group differences, not detected by the other methods.

Multiple Testing

In the conduct of medical research there is an understandable wish to find results that are ‘statistically significant’. There are many motives for this. It is intrinsically more interesting to be able to report significant differences between groups, while a failure to produce ‘significant’ results can be viewed, often incorrectly, as a failure of the study. It has been recognised for decades that there is a publication bias in the medical literature, whereby studies with ‘significant’ findings are more likely to published than equivalent studies with ‘non-significant’ findings. With academic promotions being heavily influenced by the publication record, this is a powerful reason to be able to submit papers with significant results and, as a reviewer for journals, it is apparent that many submissions are ‘spun’ to show the research in the best possible light. There may even be commercial reasons for wanting to show significant benefits of a new treatment or evaluation tool.

Historically, one of the ways that achieved ‘significant’ findings, perhaps by ignorance rather than by design, was to have a large number of outcome variables and then to make a post hoc decision as to which should be reported. If an ineffective intervention is being tested against a control group and there are 20 independent outcome measures then, on average, the trial should find one variable showing ‘significance’ at the 5% level and there is only a 36% chance that none of the variables will be ‘statistically significant’. Clearly, this makes a mockery of the concept that the significance test controls the risk of falsely rejecting the Null hypothesis. The realisation of how misleading such reporting could be was acted upon by the clinical trials community in the first instance and guidelines were implemented to ensure that a primary outcome variable was pre-specified, along with all secondary outcome variables of interest. Clinical trials are now tightly regulated but the same restrictions do not apply to the bulk of medical research.

The above example with multiple outcome variables illustrates just one of many ways in which a multiple testing situation can arise. There may be a comparison among several groups. If all pairs of groups are directly compared with each other, there is an enhanced chance that at least one ‘significant’ difference will be obtained. If observations over time are recorded and the data from each time point is analysed separately, then there is clearly more than a 5% chance of seeing at least one ‘significant’ result. In the context of this journal, an obvious example of the potential for this kind of multiple testing comes from data on the gait cycle.

This provokes the question of how the multiple testing problem can be handled. With multiple outcome variables, the clinical trials approach of identifying a primary outcome variable is to be recommended when it is applicable but it may not always be sensible in, for example, kinematic studies. If there are several groups to be compared, then some form of analysis of variance can avoid the multiple testing issue. If groups are being compared using a limited number of observations over time, the use of a mixed model can examine firstly whether there is a significant between group difference in the pattern of results over time, (a group by time interaction). If the interaction test is non-significant we can simply estimate the average between group difference, as there is no statistical justification for looking separately at each timepoint. If the patterns differ significantly, the between group differences can then be estimated at each individual timepoint, without there being a multiple testing issue, as we have used just a single test to justify different effects at different times.

For analysis of the gait cycle there are at least two ways of avoiding or alleviating the multiple testing problem. It can be alleviated by using the full gait cycle data to identify a limited number of derived parameters such as peak values. The author has recently become aware of statistical parametric mapping as a way to compare gait cycle data over the entire cycle. This method allows for the multiplicity of testing by using sophisticated multivariate statistical techniques [15]. Its implementation in Matlab (SPM, www.spm1D.org, v0.3; Matlab 2015b, Mathworks Inc.) is likely to increase the use of this technique which totally resolves the multiple testing issue. Although the author has insufficient experience with this technique to fully endorse it, this seems to offer a sound method that could be widely applicable to gait cycle data.

Although there are methods that can be used to solve the multiple testing problem in several specific circumstances, there is often no way to avoid making multiple tests of significance. The number of tests considered, even in a single table in a paper can sometimes be massive. Table 1 in the paper by Tamura et al. [16] shows 210 tests of significance. The Bonferroni correction is sometimes advocated for this multiple testing situation. This method simply multiplies the usual

p-values by the number of tests being performed, subject to a maximum value of 1. This is certainly a safe method in the sense that it ensures that the chance of a false positive finding at the 5% significance level is guaranteed to be less than 0.05. It carries a high risk, however, that important differences may be discounted as non-significant, as it is a very crude, non-discriminatory method that over-corrects in controlling the risk of a false positive finding. It is not a method that I would consider applying. On the other hand, I would not wish to avoid confronting the problem of the risk of false positive findings. My preferred method is to report unadjusted p-values but to interpret them cautiously in the light of the multiple testing that has taken place. That interpretation should, of course, take into account the number of significance tests performed and the number of 'significant' findings that might be expected even if the dataset was uninformative. It is always unacceptable to simply ignore the number of tests being performed when interpreting the results from any study.

Subgroup Analysis

If the patients in a study have some defining characteristics (e.g. age or gender), there is often interest in whether the findings in the whole sample are the same within these subgroups. The motive can be the laudable aim of extracting the maximum possible information from an expensively garnered dataset. If an intervention has not shown statistically significant benefit overall, perhaps it is effective in the elderly? If there is an overall statistically significant effect of an intervention, does this hold for both men and women. As an example, a multi-centre study of the effect of rehabilitation on gait pattern in patients with multiple sclerosis looked at overall results but also examined effects within subgroups defined by normal walking speed and by centres [17].

The obvious, but incorrect, way to conduct a subgroup analysis is to repeat the analysis used in the total sample, applying this to each subgroup in turn. Using artificial data we might obtain a difference between two 'treatments' with a mean and 95% confidence interval of 4.3 (0.1, 8.5) in males and 5.1 (-0.1, 10.3) in females. This analysis would conclude that there is a statistically significant difference in men but no statistically significant difference in women.That conclusion is semantically correct but it is neither helpful nor sensible. The observed magnitude of effect is larger in females but we are concluding no significant effect while concluding that the smaller effect in males is statistically significant. Such a paradoxical result can easily occur when the subgroups have different sample sizes and hence different statistical power.

A better approach is to ask, as a preliminary question, is the 'treatment' effect in males significantly different from the effect in females? In epidemiological jargon, we are asking if gender is an effect modifier. In statistical terms we are asking whether there is a statistically significant treatment by gender interaction. In the above artificial example the gender effects do not differ significantly and we should conclude that there is, overall, a statistically significant treatment effect but there is no evidence that the treatment effect differs between males and females.

It is necessary to think carefully before applying subgroup analyses. If there is a priori information to suggest that there may be a treatment by subgroup interaction, then it is sensible to pre-specify the interaction test as part of the analysis plan and mention this in any publication. If there is no sound reason to anticipate a subgroup effect, any subgroup analyses that are performed should be regarded as exploratory and identified as such. There is a temptation to explore all subgroup possibilities but this should be avoided. There are so many different ways in which we could form subgroups from any dataset that testing all, or even several, possibilities raises the problem of multiple testing with the accompanying risk of false positive findings. The multiple testing aspect must then be considered when interpreting the results and, in an exploratory analysis, all findings must be interpreted cautiously.

Regression Modelling

Many of the comments in this section will apply to all kinds of statistical models. These include simple linear regression, multiple linear regression, linear logistic regression, Poisson regression, generalised linear models, mixed models and many others. They also apply to analysis of variance

as that can be re-expressed as a regression. The first thing that I would urge any user of these techniques to do is to read about the assumptions that underpin the use of the particular type of model that they are employing and how these assumptions can be checked once the model has been fitted to their data.

Before anyone embarks on regression analysis they should firstly be aware of the problem of overfitting. This can occur when the number of potential explanatory variables is too large for the number of observations. The consequence is that the fitted regression model starts to reflect the random error in the data rather than the true 'population' regression model. This overfitting can produce misleading regression coefficients and associated p-values and the amount of variance explained by the regression model becomes much too large. A rule of thumb which is commonly advocated is that the number of variables fitted should be less than one-tenth of the number of observations in multiple regressions, less than one-tenth of the number of events in a survival analysis and less than one-tenth of the numbers in the numerically smaller group when undertaking multiple logistic regression. That rule of 10 is not as arbitrary as it may seem at first glance. Simulation studies have shown this to have justification in association with logistic regression [18]. In contrast, for pre-specified multiple regressions, it has been shown that a ratio of 2:1 for observations versus predictors is adequate to give a bias of less than 10% in the regression coefficients [19]. Common-sense would tell us that these rules should not be considered as set in stone but they indicate when we need to be aware of overfitting as a potential problem.

The stage of model checking is one that is commonly omitted in papers that I review for this journal and others. It is not the role of this paper to provide a compendium of how model checks can be made but to identify some of the common omissions. Numerically the most common unchecked assumption must be that of linearity. It is such a fundamental assumption of most models that it is easy to forget that it is only an assumption and not inevitable. For example, in many studies, the effect of age is something that might interfere with an association of interest and it will appear as a linear term in the final model, so that the association is said to be 'adjusted for the effect of age'. Over short age ranges that linear assumption is often reasonable and difficult to disprove but in studies containing both the young and the very old the assumption may be less reasonable. For example, Taylor [20] showed in nerve conduction studies in adults that most variables increased (or decreased) with age until reaching a maximum (minimum) in the fourth decade with a subsequent decline (increase).

Another assumption that is commonly present is that of additivity of the terms in the model. If we again consider a model that includes age, there is an assumption that the effect of age on the outcome variable will be the same whatever the values of the other variables in the model. If that assumption is false and suppose that the effect of age on the outcome variable differs according to gender, then the model needs to be changed to also include a gender by age interaction term. In epidemiological terms we would state that gender is an effect modifier for the effect of age (and vice versa). Checking the assumption of additivity is something that needs to be done with a degree of finesse. It would be possible to check all pairs of terms in a model for a significant interaction but that would introduce all of the problems of multiple testing. It will often be sensible to check just for plausible interactions or interactions that are potentially critical for the interpretation of the findings.

If the outcome variable is continuous, there will be distributional assumptions. In most cases this will be that the residuals (error terms) are Normally distributed with equal variances. As noted in a previous section, it is important to note that the assumption is **not** that the outcome variable is Normally distributed and when many variables are being fitted the residuals may conform well to normality, even if the distribution of the outcome variable is markedly non-Normal. Most statistical packages will allow checking of the residuals and this is recommended. If the assumption is clearly false then re-analysis after transformation of the outcome variable or the use of a generalised linear model may produce a more suitable model.

Elements of model checking that are often overlooked are the search for potential outliers in the data or observations with high leverage but more important is the examination of whether there are observations that are particularly influential. Influential points have the property that their omission will have a large effect on the model that is fitted. Realistically, not every paper will have

their data checked to this extent, but it is a counsel of perfection that should be aspired to and should at least be applied to papers that are expected to be pivotal in their own area. In datasets with very influential observations that might change the conclusions, a sensitivity analysis, omitting such observations, should be shown as well as the main analysis. Further details on model checking may be obtained in many statistical textbooks including a longstanding favourite by Draper and Smith [21].

A common dilemma in regression modelling is which of the variables available to us should appear in the model and which may be safely omitted. Statistical packages will usually have several automatic model selection procedures available which can make the process very simple. Typically there will be a forward stepwise procedure where the basic idea is that the most significant of the remaining variables will be added to the model in turn until all of the non-included variables are non-significant at a pre-determined level of significance (often defaulting to 0.05). A backward stepwise procedure starts with all variables in the model and removes the least significant, one at a time, until only 'significant' variables remain. These methods may be further tweaked by allowing terms to drop out of the forward stepwise procedure if they become non-significant at a subsequent stage in the process, or re-enter in the backyard stepwise procedure if they become significant. It is not widely appreciated that the forward and backwards procedures may not produce the same model for any particular dataset. When they do, it gives some reassurance about the robustness of the model. 'Best subsets regression' is another occasionally useful technique though it is not widely used.

So are these automatic methods useful? The unhelpful but honest answer is that sometimes they are but often they are contraindicated. It depends entirely on the purpose of the model, the number of observations available and the number of potential explanatory variables. In the course of reviewing, I sometimes gain the impression that stepwise methods have been used simply because they are available in the statistical package or because they have been used in other similar publications or because the authors are unaware of different approaches. I suggest that it is best practice to explain the purpose of the modelling that is being employed, and why a particular modelling method is used in preference to alternatives. This implies that, before analysis starts, the researcher should be clear about their purpose in using a regression model and always consider which methods for developing their model best suit that purpose. If they suspect that their statistical expertise may be insufficient to make the best choice, they should consider seeking statistical advice at an early stage.

Whenever regression models are used in a publication, the model selection process should be described in reasonable detail within the Methods section of the paper. If a stepwise procedure has been used, the description should include a list of all of the variables eligible for inclusion, as well as the criteria for adding or removing variables from the model. If a continuous variable such as age has also appeared in the paper in a categorised form, it should be clear whether it has been included in the model as a continuous or categorical variable. Generic descriptions of the analysis should be avoided. For example, stating that mixed models were used in analysis means practically nothing until you specify which variables were fixed, which were random and, if there were repeated observations, what covariance pattern was fitted. The guiding principle is that you should be providing sufficient detail that another researcher with access to your data would produce the same results.

Within any regression model there may be some terms that are of central interest while other terms in the model have no intrinsic interest. It may be reasonable, for reasons of space, for the tables in the paper to show only the results of main interest. It is nevertheless best practice to always include the full fitted model in the paper, even if this is only in an Appendix or as an online Supplement. As well as the fitted model, statistical packages supply a host of summary statistics, indicating the goodness of fit of the models. These are often helpful though authors should ensure they understand those measures that they present. There is a danger of 'reporting overkill' with the inclusion of statistics that are irrelevant.

A detail that should always be reported but is often omitted is the effective sample size for the model. Missing values are common in many studies and their presence can reduce the number of observations that are used in the regression modelling. For example, in multiple linear regression or multiple logistic regression, a missing value for any of the predictor variables for an individual

causes that individual to be omitted from the analysis. If using a stepwise procedure, that applies to every variable, whether or not it appears in the final model. In some circumstances, attrition can be appreciable and sometimes requires alternative modelling approaches to be used. These can sometimes be as simple as omission of variables with excessive missing values or as sophisticated as the use of multiple imputation [10].

The theme that should be coming through from these comments is the need for clarity at every step. The reader should be aware of precisely what analytical methods were used and why. They should be shown that the assumptions for the modelling were checked and were satisfied. As well as showing the 'headline' results from the analysis, the full models should be available for the reader to obtain greater detail. Any attrition from the full dataset should be apparent. How has the robustness of the findings been examined? If a paper gives the impression that some details have been 'swept under the carpet', then the credibility of the findings will, inevitably, be undermined.

Predictive modelling

There is one particular application of regression models that should be mentioned. That is its use as a method of predicting outcomes. For example, there may be interest in predicting which elderly patients will experience a fall, using clinical characteristics of the patient together with assessments of balance [22] or daily gait variables obtained from a trunk accelerometer [23]. Once a predictive model has been obtained (and these can also be obtained by machine learning methods such as classification and regression tree analysis), there is the question as to how the success of the predictive model should be evaluated. There are many metrics used for this purpose including the sensitivity, specificity, accuracy and area under the receiver-operating characteristic curve (C statistic) [24]. Whatever metric(s) is/are chosen, in practice, most papers proposing predictive models evaluate them on the same dataset that has been used to generate the model. This highly circular approach means that there is a favourable bias on every metric that can be chosen. Furthermore, that bias is even greater when the model is overfitted, as is often the case [25] and is usually the situation with machine learning methods that implicitly use far more explanatory variables than regression based methods. The ideal for predictive models is to have a training dataset that generates the model and an independent evaluation dataset to assess the properties of the model. With this in mind, researchers often split a single dataset into training and evaluation subsets, often in the ratio of 2:1. This avoids the problem of biased metrics but it can be seen to be inefficient (unless sample sizes are huge) as the sample size for prediction is reduced.

There are two main methods that use the full dataset for prediction but attempt to correct for the over-optimistic metrics that are obtained when the same dataset is used for validation. These are cross-validation and bootstrapping. There are many varieties of cross-validation of which k-fold cross-validation is perhaps the most popular. The dataset is randomly divided into k distinct equal sized subgroups and the modelling strategy is applied to the data from k-1 of them to obtain a predictor that is tested on the remaining subgroup to obtain the desired metrics. This is repeated k times and the results are averaged. The entire process may be repeated several times for greater stability in the results and we can have, for example a 10 x10% cross-validation where k=10 and the process is repeated 10 times [26].

In bootstrapping, if the sample size is n, samples of size n are drawn with replacement from the original sample. For each sample the modelling strategy is used to form a predictor which is then evaluated on the same sample that generated it and on the original sample. The difference estimates the 'over-optimism' or bias. These differences are averaged from a large number of samples to give a measure of the bias in the method and corrected metrics can be generated. In simulation studies based on one clinical dataset, Steyerberg et al. [26] found that bootstrapping performed better than alternatives. More recently, Smith et al. [27] have undertaken studies based on several clinical datasets and, using the C-statistic as their metric, also marginally preferred the bootstrap method. They warned though that all methods showed bias when models were overfitted or when the value of C was small, with bootstrapping giving optimistic estimates while cross-validation gave pessimistic estimates.

All studies have shown that the naive method of evaluation is to be avoided at all costs. It is therefore recommended that any predictive model should always be evaluated using either cross-validation or bootstrapping. The results from Smith et al. [27] indicate that least 200 bootstrap samples or 20 replications of 10-fold cross-validation should be used. Precise details of the methods used should always be included in any submission as the author's experience in reviewing has shown that it is very easy for them to be mis-applied.

Intraclass Correlation Coefficients, Levels of Agreement and Variance Components

Levels of agreement are sought in many contexts. Bishop, Hillier and Thewlis [28] examined the intra-rater agreement and the inter-rater agreement of the Adelaide in-shoe foot model in two separate series of patients. Wickstrom, Stergiou and Kyvelidou [29] examined the reliability of centre of pressure measures in infants through the stages of sitting. Cheng et al. [30] examined the intra- and inter-observer reliability of radiographic measurements as part of a study on compensatory mechanisms in the pelvis and lower extremities in patients with pelvic incidence and lumbar lordosis mismatch. Saner, Washabaugh and Krishnan [31] examined hip and knee kinematic variables using low-cost webcam technology to determine agreement on two different days. These happened to be the first four papers published in volume 56 of Gait and Posture and all used intraclass correlation coefficients (ICC) among other measures of reliability. Therefore it is a topical subject. I am sure that authors of papers feel there is pressure to report ICCs as their presence is so widespread. The ICCs are then usually interpreted qualitatively using criteria such as those of Rosner [32]. Sadly, for the readership of these papers, the use of ICCs can be positively misleading. It is the belief of this author that ICCs rarely have a place in studies in which different measurement methods are compared, where inter- and intra-rater variability is assessed or in more general situations where components of variability in measurement are being assessed.

The problem with ICCs is that they are strongly dependent on the between subject variability. If two studies are undertaken with identical inter-and intra-rater variability, it is perfectly possible that a study undertaken on a highly heterogeneous sample of subjects will be assessed by ICCs as having excellent reproducibility whereas an otherwise identical study on a closely homogeneous sample of subjects will be assessed as having poor reproducibility. This is exemplified in Table 2 where artificial data have been constructed to examine ICCs, Standard Error of Measurement (SEM) and components of variance in two samples. In one of these, the subjects are quite similar as might happen if the subjects are of the same gender and age and with normal gait. The second sample mimics the choice of subjects of both genders with a wide age range and including both subjects with normal and abnormal gait. The data has been deliberately constructed so that the differences between the pairs of observations are identical in both samples and yet the ICCs differ substantially. Conventionally, the first sample would be said to show poor reproducibility and the second excellent reproducibility, despite the identical SEMs.

It is not the purpose of this paper to attempt to provide a template for the design and analysis of those studies where ICCs are currently used. For some studies the approach of Bland and Altman [33] will be suitable and every researcher in this area should be aware of this work. Perhaps the most informative studies and analyses are those where as many sources of variability as possible are examined and quantified. Apart from the direct knowledge of the extent of various sources of variability, such knowledge can be used to plan more efficient future studies or improve the assessments of patients so that meaningful changes in measurements can be detected. Chia and Sangeux [34] show how to estimate five sources of variability from gait analysis in their example and discuss how these variance components can be of clinical use. It is estimation of these variance components that should be central in reliability studies. Chia and Sangeux sum things up nicely in their conclusions where they state "We believe the variance components should always be the fundamental quantities to be computed when investigating issues related to variability, reliability or repeatability. This is because the variance components are actually variances which are well-understood statistical quantities in meaningful units … they are applicable to a wide range of scenarios".

2x2 Contingency Tables

This is such a simple data structure that one would expect its analysis to be simple and uncontroversial. If we characterise the data by membership of either group A or group B and have a binary outcome that we will label success or failure, then the data is completely summarised by the number of successes in group A, the number of failures in group A, the number of successes in group B and the number of failures in group B. We will later refer to the probabilities of success as P(A) and P(B).

Despite this simplicity of structure, there have been numerous methods proposed to test the statistical significance of the association between group and outcome. There are also several different ways in which the magnitude of the difference in outcomes between the two groups can be estimated and 95% confidence limits calculated. Among statisticians there have been sometimes acrimonious discussions which Martin Bland [35] has described as “generating almost as much heat as light”.

Initially we will look at ways to conduct a test of significance for a 2x2 contingency table.

Significance Testing

We will take as an example data from Bernard et al. [36] from the trauma literature. In adult patients with severe brain injury the use of urban road ambulance-based paramedic drug-assisted rapid sequence intubation (RSI) was compared in a randomised controlled trial to transport and subsequent intubation by a physician in the hospital emergency department. One of the secondary outcomes was a good neurological outcome at 6 months, as shown in Table 3. The table also shows the results of applying various tests of significance to that 2x2 contingency table. It is apparent that different conclusions with respect to significance/non-significance at the conventional 5% level would arise dependent on the test applied. The Annals of Surgery accepted the p-value of 0.046 but other referees, including myself, would come to a different conclusion. This is an unsatisfactory situation and in this paper it is proposed that one particular test should become the standard method for acceptance in the medical research journals.

Firstly though, a brief recap of why there are a variety of tests might aid clarity without becoming bogged down in too much detail. (A comprehensive coverage of the 2x2 table would need a small textbook). There are two key features that distinguish the analysis of 2x2 tables from other well-known statistical tests such as the Student t-test. As the data are discrete, for any particular dataset, the p-values from any test can only take on a limited number of possible values. This means that, in general, it is not possible to have exactly a 5% chance of rejecting the Null hypothesis when it is true, as can happen with the t-test. Some of the contender tests are advocated because they can get closer to the nominal 5% level than other tests and hence have a greater statistical power to detect differences between the groups as statistically significant. Increased power is a desirable property but there can be a downside to some of these more powerful techniques, as we shall see shortly. The second feature of the analysis is that we don’t automatically have symmetry in our comparisons in the way we do with the t-test. The possible values that can occur for P(A) - P(B) are not identical to the values that can occur for P(B) -P(A), unless we happen to have identical numbers in the two groups or an equal overall number of successes and failures. The way in which that lack of symmetry is handled provides additional possibilities for the method of analysis and for the conclusions that we can obtain.

Before presenting specific tests for 2x2 tables, it will be helpful to review how tests of significance are applied in general and in medical research in particular. Traditionally, tests of significance are presented in terms of setting a Null hypothesis, calculating a suitable test statistic and then evaluating the probability of obtaining values of the test statistic as or more extreme than the one observed. Conventionally, if that probability is greater than 0.05 we accept the Null hypothesis and otherwise we reject the Null hypothesis. That is, we come to a binary decision about acceptance or rejection of the Null hypothesis. In medical applications when we are comparing two groups we are, de facto, going one stage further. When applying a t-test, for example, we will have a tertiary divide rather than a binary divide. Having found that the t-test gives a statistically significant difference between the groups we will not simply report the test in that way. We will

either conclude that the mean in group A is significantly higher than the mean in group B or we will conclude that the mean in group B is higher than the mean in group A. In effect, we are conducting two one-tailed tests at the 2.5% significance level, even if we do not explicitly think of it in these terms. We don't need to think of it explicitly in this way because the test is symmetrical and showing overall significance at the 5% level is equivalent to having performed these two one-tailed tests. When we have a 2x2 contingency table, the lack of symmetry in P(A) - P(B) and P(B) - P(A) means that we will only be able to make that desirable tertiary divide if we explicitly set out to perform two one-tailed tests at the 2.5% significance level. That consideration will lead to our later recommendations.

We will now go through some of the more common tests that have been proposed.

Fisher's Exact Test

This test is based on calculating, under the Null hypothesis of no association, the probabilities of every possible 2x2 table that could give rise to the the observed row and column totals (e.g. in Table 3, the number in the RSI group, the number in the control group, the total number of good neurological outcomes and the total number of bad neurological outcomes). There are fewer possible 2x2 tables than might appear at first glance since if we know, for example, the number of good outcomes in the RSI group, then the other entries in the table can be obtained by subtraction from the row or column totals. In Table 3 the observed probability of a good outcome is higher in the RSI group (80/136). From this we can calculate a one-tailed exact probability, under the Null hypothesis, by adding together the individual probabilities of the tables where the number of good outcomes in the RSI group is 80 or higher. This one-tailed exact probability is shown in virtually every statistical package.

For a symmetrical test such as the t-test, we obtain a two-tailed test of significance by simply doubling the one-tailed probability, under the Null hypothesis. This can also be done to create a two-tailed test for 2x2 tables [37]. We simply take the one-tailed exact probability and report double its value as the p-value. The doubling ensures that symmetry is maintained and that is the test that I recommend. It means that we have an equal chance of rejecting the Null hypothesis in both directions from the Null. It also guarantees that if we reject the Null hypothesis at the conventional 5% level, we can either conclude that P(A) is significantly higher than P(B) or that P(B) is significantly higher that P(A) at the 2.5% significance level conventionally used for superiority. In the example in Table 3 we conclude that p=0.06 and the two treatment groups do not differ significantly at the 5% level.

A different and more widely used alternative way of generating a two-tailed test of significance is output as the exact two-tailed probability in statistical packages. This takes the one-tailed exact probability and adds to it the probability of every contingency table in the opposite direction that has a probability less than or equal to the probability of the observed table. This is attractive as it usually results in a lower p-value with a corresponding gain in power, since it allows observed p-values to come closer to the conventional 5% level. This undeniable increase in power comes with a cost in terms of the inferences that can be drawn. In our example in Table 3, the two-tailed exact probability of 0.046 allows us to reject the Null hypothesis of no association between the groups and neurological outcome. There is a logical inconsistency though, as we do not have enough evidence to conclude that RSI is superior to the control group in terms of neurological outcome, as the one-tailed exact p-value is 0.03, nor of course can we conclude that the control group is superior. Thus we are rejecting the Null hypothesis without being able to conclude either that P(A) > P(B) or P(B) >P(A). That kind of paradoxical conclusion is unsatisfactory in the view of the author and, for that reason, this version of Fisher's exact test should be avoided in medical research.

There is yet another even more powerful version of Fisher's exact that is occasionally used, though its absence from many statistical packages has doubtless limited its use. This tests allows for the granularity of the data and hence the observable p-values by taking the two-tailed exact probability and subtracting half of the probability of the observed configuration. This is known as the mid-P method. It has the advantage of still greater power but has the same drawback of possibly paradoxical conclusions.

In summary, the recommendation from this paper is to apply the symmetrical version of Fisher's exact test where the p-value is reported as double the one-tailed exact probability.

The Chi-squared Test

For tables with more rows or columns than the 2x2 table the general form of the chi-squared test is uniquely defined. In the special case of a 2x2 table there are two possibilities. As well as the standard form of the test there is also a continuity corrected version, usually described as Yates' correction for continuity [38]. There are different ways in which these two tests can be theoretically derived. For the purpose of this paper we will consider, particularly, the chi-squared test as an approximation to Fisher's exact test. As sample sizes increase, the exact two-tailed probability becomes better and better approximated by the uncorrected form of the chi-squared test. Similarly, as the sample sizes increase, the symmetrical form of Fisher's exact test is better and better approximated by the chi-squared test with Yates' correction for continuity.

From the arguments of the previous section, it is therefore logical to use the continuity adjusted test rather than the uncorrected version. Until quite recently, the use of the chi-squared test would have been preferred to the use of Fisher's exact test because it is so much simpler to calculate. Modern computing power and the widespread availability of statistical packages negate this advantage as even numbers such as those in Table 3 can be analysed using Fisher's exact test almost instantaneously.

Summary and Recommendation for Significance Tests on 2x2 Tables

Tests of significance on 2x2 tables should routinely be made using Fisher's exact test, reporting the p-value as double the one-tailed exact probability. Modern computing power and software make the use of the chi-squared test unnecessary in most instances but if the test has to be used, Yates' correction for continuity should always be employed.

Estimation and Confidence Intervals

Estimates of 'treatment' effects from 2x2 tables can be made in three main ways. The absolute difference in 'success rates' can be used directly. The relative effects in the two groups can be estimated using odds ratios or relative risks. The method of choice will depend on the particular research questions that are being posed and there is no 'best' summary measure. Whatever measure is used, confidence intervals for the chosen parameter should be presented and, conventionally, 95% confidence intervals are usually chosen. The confidence intervals that are most commonly used in software packages are based on large sample (asymptotic) theory. In most instances these will give sufficient accuracy but care needs to be exercised when samples are relatively small. If the statistical software offers test-based or 'exact' confidence limits, these are to be preferred. Particular care should be taken when the p-value is close to 0.05 that the 95% confidence limits are compatible with the reported p-value.

Table 3 shows the estimates and 95% confidence intervals for the data from Bernard et al. [36]. For the absolute difference in success rates, the first confidence interval shown is based on the 'standard' asymptotic approach and corresponds to the uncorrected form of the chi-squared test. The 95% confidence interval does not contain 0, in line with p=0.046 from the chi-squared test. The second confidence interval is adjusted for continuity and this confidence interval does include 0, corresponding to the p-value of 0.06 when Yates' continuity adjusted chi-squared was calculated. Therefore, for concordance with the recommendation on significance testing, the latter set of confidence limits should routinely be applied. These can readily be obtained using the free online calculator from VassarStats [39]. Only with very small sample sizes is this likely to be unreliable.

Odds ratios are the most commonly quoted summary measure and the two sets of confidence intervals shown in Table 3 correspond to the asymptotic limits and the exact limits respectively. As before, the asymptotic limits exclude the Null value (of 1) while the exact limits include this value. The asymptotic limits are readily available in virtually all statistical packages and in online sources such as VassarStats [39]. The exact limits were obtained using SAS® PROC FREQ and will be available in some, but not all, statistical software. In this instance it would be important to show

the exact confidence limits but, in less critical circumstances, the asymptotic limits would be adequate, if not quite the ideal.

Relative risks, also known as rate ratios or risk ratios, are less commonly employed but asymptotic limits are readily available when these are used. These are shown for our example in Table 3. Attempts to obtain exact limits in SAS® were unsuccessful as nonsense values were obtained.

In summary, if absolute differences in proportions are being calculated, confidence limits should either use an exact method or a continuity adjusted method. For odds ratios or risk ratios, the 'standard' confidence intervals will usually be adequate but with small sample sizes or p-values close to 0.05, exact methods are needed.

Summary and Recommendations

Authors of research publications often face pressure on space and this may be one reason why statistical methods are often reported in insufficient detail. In the past, that would have been a reasonable explanation but the increasing possibility of supplying additional information in online supplements means there is now little excuse for failing to present the statistical analysis with clarity. The call for clarity in presentation is an underlying theme throughout this paper. The target should always be to present an analysis in sufficient detail that another researcher could produce the same results from the same data. I would go further and suggest that, whenever there are alternative methods that could have been used, the authors should justify the approach they have taken.

The recommendations that have been made in this paper are summarised in Table 4. Of note, is the fact that several of these recommendations argue against the use of methods that some might regard as 'standard'. The use of post-hoc power, intra-class correlation coefficients and the uncorrected chi-squared test for 2x2 contingency tables are cases in point. All too commonly there can be fashions in research, facilitated by what is available in statistical software, and the reporting may be adapted to follow these fashions without questioning the basics. I would urge researchers to take a step backwards from a semi-automatic data analysis and think deeply about what they are trying to achieve. If they lack certainty about the methods they are considering, I would urge consulting with an experienced medical statistician at the earliest possible stage.

References

[1] Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. Br Med J. 1977; i: 85-87.

[2] Fernandes-Taylor S, Hyun JK, Reeder RN, Harris AH. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. BMC Research Notes. 2011;4:304.

[3] Prescott RJ, Civil I. Lies, damn lies and statistics: errors and omission in papers submitted to Injury 2010-2012. Injury. 2013;44:6-11.

[4] Bland M. An Introduction to Medical Statistics (4th ed). Oxford, Oxford University Press; 2015.

[5] Kirkwood BR and Sterne JAC. Essential Medical Statistics (2nd ed). Oxford, Blackwell Science; 2003.

[6] Altman DG. Practical Statistics for Medical Research. London, Chapman and Hall; 1991.

[7] Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Melbourne D, Egger M, Altman DG. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340 :c869

[8] Bishop C, Paul G, Thewlis D. Recommendations for the reporting of foot and ankle models. J Biomech. 2012; 45: 2185-2194.

[9] Lang TA, Altman DG. Basic Statistical Reporting for Articles Published in Biomedical Journals: The "Statistical Analyses and Methods in the Published Literature" or The SAMPL Guidelines" in: Smart P, Maisonneuve H, Polderman A (eds). Science Editors' Handbook, European Association of Science Editors, 2013.

[10] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338:b2393.

[11] Nelder JA, Wedderburn RWM. Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General) 1972; 135(3): 370-384.

[12] Fagerland MW, t-tests, non-parametric tests, and large studies—a paradox of statistical practice? BMC Medical Research Methodology 2012; 12:78.

[13] Brown H, Prescott R. Applied Mixed Models in Medicine (3rd ed). Chichester, John Wiley & Sons; 2015.

[14] Stewart S, Pearson J, Rome K, Dalbeth N, Vandal AC. Analysis of data collected from right and left limbs: accounting for dependence and improving statistical efficiency in musculoskeletal research. Gait Posture 2018; 59: 182-187.

[15] Pataky TC, Robinson MA, Vanrenterghem J. Vector field statistical analysis of kinematic and force trajectories. J Biomech. 2013; 46: 2394-401.

[16] Tamura K, Radzak KN, Vogelpohl RE, Wisthoff BA, Oba Y, Hetzler RK, Stickley CD. The effects of ankle braces and taping on lower extremity running kinematics and energy expenditure in healthy, non-injured adults. Gait Posture. 2017; 58: 108-114.

[17] Leone C, Kalron A, Smedal T, Normann B, Wens I, Eijnde BO, Feys P. Effects of Rehabilitation on Gait Pattern at Usual and Fast Speed Depend on Walking Impairment Level in Multiple Sclerosis. Int J MS Care. 2018; in press https://doi.org/ 10.7224/1537-2073.2015-078

[18] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996; 49:1373-9.

[19] Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. J Clin Epidemiol. 2015; 68: 627-636.

[20] Taylor PK, Non-linear effects of age on nerve conduction in adults. J Neurol Sci. 1984; 66: 223-234.

[21] Draper NR, Smith H. Applied Regression Analysis (3rd ed). Chichester, John Wiley & Sons; 1998.

[22] Shumway-Cook A, Baldwin M, Polissar NL, Gruber W. Predicting the Probability for Falls in Community-Dwelling Older Adults. Physical Therapy. 1997; 77: 812–819.

[23] van Schooten KS, Pijnappels M, Rispens SM, Elders PJM, Lips P, van Dieën JH. Ambulatory Fall-Risk Assessment: Amount and Quality of Daily-Life Gait Predict Falls in Older Adults. J. Gerontol: Series A 2015: 70: 608–615.

[24] Akosa JS. Paper 942-2017:Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data. support.sas.com/resources/papers/proceedings17/0942-2017.pdf

[25] Shany T, Wang K, Liu Y, Lovell NH, Redmond SJ. Review: Are we stumbling in our quest to find the best predictor? Over-optimism in sensor-based models for predicting falls in older adults. Healthc Technol Lett. 2015; 2: 79–88.

[26] Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001; 54:774-81.

[27] Smith GCS, Seaman SR, Wood AM, Royston P, White IR. Correcting for Optimistic Prediction in Small Data Sets. Am J Epidemiol. 2014; 180: 318–324.

[28] Bishop C, Hillier S, Thewlis D. The reliability of the Adelaide in-shoe foot model. Gait Posture 2017; 56: 1-7.

[29] Wichstrom J, Stergiou N, Kyvelidou A. Reliability of centre of pressure measurements for assessing the development of sitting postural control through the stages of sitting. Gait Posture 2017; 56: 8-13.

[30] Cheng X, Zhang K, Sun X, Zhao C, Li H, Zhao J. Analysis of compensatory mechanisms in the pelvis and lower extremities in patients with pelvic incidence and lumbar lordosis mismatch. Gait Posture 2017; 56: 14-18.

[31] Saner RJ, Washabaugh EP, Krishnan C. Relaible sagittal plane kinematic gait assessments are feasible using low-cost webcam technology. Gait Posture 2017; 56: 19-23.

[32] Rosner B. Fundamentals of Biostatistics (5th ed). New York, Duxbury Thomson Learning; 2000.

[33] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; i: 307–310.

[34] Chia K, Sangeux M. Quantifying sources of variability in gait analysis. Gait Posture 2017; 56: 68-75.

[35] Bland M. An Introduction to Medical Statistics (4th ed). Oxford, Oxford University Press; 2015

[36] Bernard SA, Nguyen V, Cameron P, Masci K, Fitzgerald M, Cooper DJ, Walker T, Std BP, Myles P, Murray L, David, Taylor, Smith K, Patrick I, Edington J, Bacon A, Rosenfeld JV, Judson R. Prehospital rapid sequence intubation improves functional outcome for patients with severe traumatic brain injury: a randomized controlled trial. Ann Surg. 2010; 252: 959-65.

[37] Yates, F. Tests of significance for 2 x 2 contingency tables (with discussion) J. R. Statist. Soc. A, 1984; 147, 426-463.

[38] Yates F. Contingency table involving small numbers and the χ^2 test. J Roy Stat Soc Supp 1934; **1**: 217–235.

[39] VassarStats: Website for Statistical Computation. vassasstats.net

Table 1: Examples of Reporting Guidelines for different subjects

Subject		Reporting Guideline
Basic statistical reporting in biomedical journals		SAMPL
Clinical trials		CONSORT
Systematic reviews		PRISMA
Observational studies		STROBE
Studies of diagnostic accuracy		STARD
Prediction models		TRIPOD

Table 2: Artificial data to illustrate changes in intra-class correlations according the amount of between subject variation in datasets with identical differences

	Hypothetical Dataset 1			Hypothetical Dataset 2	
Subject	Test 1	Test 2		Test 1	Test 2
1	23.2	26.0		18.2	21.0
2	23.4	23.0		16.4	16.0
3	23.5	23.3		21.5	21.3
4	23.7	25.3		23.7	25.3
5	23.8	24.3		20.0	20.5
6	24.0	24.9		19.6	20.5
7	24.2	23.2		24.8	23.8
8	24.4	26.4		22.2	24.2
9	24.6	24.6		20.3	20.3
10	24.8	25.0		14.1	14.3
	Components of variance			Components of variance	
	Subjects=0.1667			Subjects=9.7960	
	Occasions=0.0659			Occasions=0.0659	
	Residual=0.7991			Residual=0.7991	
	ICC=0.162			ICC=0.919	
	SEM=0.93			SEM=0.93	

Table 3: Application of different significance tests and confidence intervals to data on neurological outcome from Bernard et al [36]

	Rapid Sequence Intubation Group	Hospital Intubation Group	Total
Good neurological outcome	80 (51.0%)	56 (39.4%)	136
Poor neurological outcome	77 (49.0%)	86 (60.6%)	163
Total	157	142	299
	Tests of Significance		
	$\mathbf{X}^2$=3.99, p=0.046		
	$\mathbf{X}_c^2$=3.54, p=0.060		
	Fisher's two-tailed exact test: p=0.049		
	2 x Fisher's one-tailed exact test: p=0.060		
	Summary measures (95% confidence limits)		
	Difference in success rates: 11.5% (0.2%, 22.4%)[a] (-0.3%, 22.8%)[b]		
	Odds ratio: 1.60 (1.01, 2.53)[a] (0.98, 2.59)[c]		
	Rate ratio: 1.29 (1.00, 1.67)[a]		

a) asymptotic 95% confidence limits
b) continuity adjusted 95% confidence limits
c) exact 95% confidence limits

values for a) and b) computed using VassarStats: Website for Statistical Computation (vassarstats.net) and c) using SAS® PROC FREQ

Table 4: Summary of main recommendations

Reporting Guidelines:
• Use available guidelines as appropriate.

Missing Values:
• Be totally transparent in reporting missing values.
• Consider using multiple imputation.

Parametric and Non-parametric Tests:
• Analyse using parametric methods on most occasions.
• Transform data when necessary to improve symmetry of the distribution.
• Non-parametric methods should be considered with smaller sample sizes when symmetrical distributions cannot be obtained.

Testing for Normality:
• Formal statistical tests of significance are never recommended but graphical methods can be helpful.

Tests of Significance, Confidence Intervals and Post-hoc Power:
• Use 95% confidence intervals to complement significance tests for main outcomes.
• Do not compute post-hoc power.

Effect Size
• Effect sizes are not generally required but may be useful in particular circumstances.

Sample Size:
• Formal sample size calculations should be based on what is clinically important.
• The sample sizes used in a study should always be justified, even when not based on a formal statistical calculation.

Independence of Observations:
• If observations are not independent, analyse using appropriate statistical methods such as mixed (multi-level) models.

Multiple Testing:
• Use methods to reduce the number of tests performed if this is possible (see text for examples).
• In most circumstances, avoid the use of Bonferroni corrections.
• Be aware that multiple testing increases the risk of false positive findings and interpret accordingly.

Subgroup Analysis
• Subgroup analyses should be limited in number.
• Such analyses should be regarded as exploratory, unless pre-specified to confirm a hypothesis.
• Analysis should always be based on the significance of subgroup by intervention interactions.

Regression Modelling:
• Be aware of the assumptions underpinning the model, check that these assumptions are satisfied and report the checks that have been made.
• Avoid ‘overfitting’ models with too many variables for the number of observations.
• Explain the reasons for the modelling strategy that has been used and describe it in detail.
• Present the full models fitted with regression coefficients, standard errors etc. even if only in an Appendix.
• Always report the number of subjects actually used in fitting the model.

Predictive modelling:
• When validating predictive models, use a fully independent dataset when available, or use bootstrapping or cross-validation. Validation using the same dataset is contraindicated.

Intraclass Correlation Coefficients, Level of Agreement and Variance Components:
• Do not use ICCs to assess levels of agreement.
• Use measures such as standard error of measurement or minimum detectable change that are directly clinically relevant to assess repeatability and levels of agreement.
• Employ study designs that allow assessment of multiple causes of variability and estimate and report these causes using components of variance.

2x2 Contingency Tables:
• Tests of significance on 2x2 tables should routinely be made using Fisher's exact test, reporting the p-value as double the one-tailed exact probability.
• Confidence intervals for the difference in two proportions should use a method that employs a continuity correction or an 'exact' method, to obtain results compatible with the significance test.
• For odds ratios or risk ratios, the 'standard' confidence intervals will usually be adequate but with small sample sizes or p-values close to 0.05, exact methods are needed.